

Characterization of the equivalence of robustification and regularization in linear, median, and matrix regression

Dimitris Bertsimas* Martin S. Copenhaver†

November 25, 2014

Abstract

Sparsity is a key driver in modern statistical problems, from linear regression via the Lasso to matrix regression with nuclear norm penalties in matrix completion and beyond. In stark contrast to sparsity motivations for such problems, it is known in the field of robust optimization that a variety of vector regression problems, such as Lasso which appears as a loss function plus a regularization penalty, can arise by simply immunizing a nominal problem (with only a loss function) to uncertainty in the data. Such a *robustification* offers an explanation for why some linear regression methods perform well in the face of noise, even when these methods do not produce reliably sparse solutions. In this paper we deepen and extend the understanding of the connection between robustification and regularization in regression problems. Specifically,

- (a) In the context of linear regression, we characterize under which conditions on the model of uncertainty used and on the loss function penalties robustification and regularization are equivalent.
- (b) We show how to tractably robustify median regression problems.
- (c) We extend the characterization of robustification and regularization to matrix regression problems (matrix completion and Principal Component Analysis).

1 Introduction

The modern data deluge has led to a fundamental re-envisioning of the possibilities for statistical regression. One common theme underlying many of these methods, especially Lasso [30] and nuclear norm minimization [20], is sparsity. The modern ubiquity of such sparsity-driven methods is propelled by both theoretical results underpinning their utility (e.g., vis-à-vis compressed sensing [12, 16, 19, 1, 14]) and, perhaps more importantly, their practical scalability due to recent advances in convex optimization [11].

In contrast to sparsity-driven motivations for a variety of statistical regression problems, in the context of linear regression earlier work [21, 31, 2] has shown that under certain conditions the regularized problems result from the need to immunize, or *robustify*, the statistical problem against noise in the data.

*Boeing Professor of Operations Research, Co-Director, Operations Research Center, Massachusetts Institute of Technology. Email: dbertsim@mit.edu.

†Operations Research Center, Massachusetts Institute of Technology. Email: mcopen@mit.edu. Supported by Department of Defense, Office of Naval Research, through the National Defense Science and Engineering Graduate (NDSEG) Fellowship.

Our goal in this paper is to shed new light on the relationship between robustification and regularization for linear, median, and matrix regression problems. Specifically, our contributions include:

1. In the context of linear regression we demonstrate that in general such a robustification procedure is not equivalent to regularization. We characterize under which conditions on the model of uncertainty used and on the loss function penalties one has that robustification is equivalent to regularization. As a result we achieve a deeper understanding of what regularization accomplishes.
2. Consistent with work connecting robustification to Lasso [31], we emphasize that regularization does not lead to sparsity, as is widely believed. While Lasso only leads to provably sparse solutions under appropriate coherence assumptions, such as under the well-known Restricted Isometry Property, it can *always* be seen as a robustification of a nominal regression problem under an appropriate model of uncertainty. We contend that the success of Lasso is primarily due to its robustness properties. This is consistent with recent work in [6] on provably sparse methods, which perform better than Lasso.
3. We extend the proposed framework to median regression and we demonstrate how robustifying median regression in problems leads to corresponding regularized problems, which we solve by the mixed integer optimization methods presented in [8].
4. We carry out our analysis in the matrix completion and Principal Component Analysis (PCA) settings. We characterize under which conditions on the model of uncertainty there is equivalence of robustification and regularization.

The structure of the paper is as follows. In Section 2, we review background on norms and consider robustification and regularization in the context of linear regression, focusing both on their equivalence and non-equivalence. In Section 3, we study robustification for median regression. In Section 4, we turn our attention to regression with underlying matrix variables, considering in depth both matrix completion and PCA. In Section 5, we include some concluding remarks.

2 A robust perspective of linear regression

2.1 Norms and their duals

In this section, we introduce the necessary background on norms which we will use to address the equivalence of robustification and regularization in the context of linear regression. Given a vector space $V \subseteq \mathbb{R}^n$ we say that $\|\cdot\| : V \rightarrow \mathbb{R}$ is a *norm* if for all $\mathbf{v}, \mathbf{w} \in V$ and $\alpha \in \mathbb{R}$

1. If $\|\mathbf{v}\| = 0$, then $\mathbf{v} = 0$,
2. $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$ (absolute homogeneity), and
3. $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ (triangle inequality).

If $\|\cdot\|$ satisfies conditions 2 and 3, but not 1, we call it a *seminorm*. For a norm $\|\cdot\|$ on \mathbb{R}^n we define its dual, denoted $\|\cdot\|_*$, to be

$$\|\boldsymbol{\beta}\|_* := \max_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}'\boldsymbol{\beta}}{\|\mathbf{x}\|},$$

where \mathbf{x}' denotes the transpose of \mathbf{x} (and therefore $\mathbf{x}'\boldsymbol{\beta}$ is the usual inner product). For example, the ℓ_p norms $\|\boldsymbol{\beta}\|_p := (\sum_i |\beta_i|^p)^{1/p}$ for $p \in [1, \infty]$ satisfy a well-known duality relation: ℓ_p is dual to ℓ_{p^*} , where $p^* \in [1, \infty]$ with $1/p + 1/p^* = 1$. More generally for matrix norms $\|\cdot\|$ on $\mathbb{R}^{m \times n}$ the dual is defined analogously:

$$\|\boldsymbol{\Delta}\|_* := \max_{\mathbf{A} \in \mathbb{R}^{m \times n}} \frac{\langle \mathbf{A}, \boldsymbol{\Delta} \rangle}{\|\mathbf{A}\|},$$

where $\langle \cdot, \cdot \rangle$ denotes the trace inner product: $\langle \mathbf{A}, \boldsymbol{\Delta} \rangle = \text{Tr}(\mathbf{A}'\boldsymbol{\Delta})$, where \mathbf{A}' denotes the transpose of \mathbf{A} . We note that the dual of the dual norm is the original norm [11]. Throughout the paper, when several arbitrary norms or seminorms appear together we also use g and h to denote (semi)norms in addition to the usual $\|\cdot\|$ notation.

Three widely used choices matrix norms (see [23]) are Frobenius, spectral, and induced norms. The definitions for these norms are given below for $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$ and summarized in Table 1 for convenient reference.

1. The p -Frobenius norm, denoted $\|\cdot\|_{F_p}$, is the entrywise ℓ_p norm on the entries of $\boldsymbol{\Delta}$:

$$\|\boldsymbol{\Delta}\|_{F_p} := \left(\sum_{ij} |\Delta_{ij}|^p \right)^{1/p}.$$

Analogous to before, F_p is dual to F_{p^*} , where $1/p + 1/p^* = 1$.

2. The p -spectral (Schatten) norm, denoted $\|\cdot\|_{\sigma_p}$, is the ℓ_p norm on the singular values of the matrix $\boldsymbol{\Delta}$:

$$\|\boldsymbol{\Delta}\|_{\sigma_p} := \|\boldsymbol{\mu}(\boldsymbol{\Delta})\|_p,$$

where $\boldsymbol{\mu}(\boldsymbol{\Delta})$ denotes the vector containing the singular values of $\boldsymbol{\Delta}$. Again, σ_p is dual to σ_{p^*} .

3. Finally we consider the class of induced norms. If $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}$ are norms, then we define the induced norm $\|\cdot\|_{(h,g)}$ as

$$\|\boldsymbol{\Delta}\|_{(h,g)} := \max_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{g(\boldsymbol{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}.$$

An important special case occurs when $g = \ell_p$ and $h = \ell_q$. When such norms are used, (q, p) is used as shorthand to denote (ℓ_q, ℓ_p) .

Name	Notation	Definition	Description
p -Frobenius	F_p	$\left(\sum_{ij} \Delta_{ij} ^p \right)^{1/p}$	entrywise ℓ_p norm
p -spectral (Schatten)	σ_p	$\ \boldsymbol{\mu}(\boldsymbol{\Delta})\ _p$	ℓ_p norm on the singular values
Induced	(h, g)	$\max_{\boldsymbol{\beta}} \frac{g(\boldsymbol{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}$	induced by norms g, h

Table 1: Matrix norms on $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$.

2.2 Uncertain regression

We now turn our attention to uncertain linear regression problems and regularization. The starting point for our discussion is the standard problem

$$\min_{\beta \in \mathbb{R}^n} g(\mathbf{y} - \mathbf{X}\beta),$$

where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{X} \in \mathbb{R}^{m \times n}$ are data and g is some convex function, typically a norm. For example, $g = \ell_2$ is least squares, while $g = \ell_1$ is known as least absolute deviation (LAD). In favor of models which mitigate the effects of overfitting these are often replaced by the regularization problem

$$\min_{\beta} g(\mathbf{y} - \mathbf{X}\beta) + h(\beta),$$

where $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is some function, typically taken to be convex. This approach often aims to address overfitting by penalizing the complexity of the model, measured as $h(\beta)$. For example, taking $g = \ell_2^2$ and $h = \ell_2^2$, we recover the so-called regularized least squares (RLS), also known as ridge regression [22]. The choice of $g = \ell_2^2$ and $h = \ell_1$ leads to Lasso, or least absolute shrinkage and selection operator, introduced in the seminal work of [30]. Lasso is often employed in scenarios where the solution β is desired to be sparse, i.e., β has very few nonzero entries.

In contrast to this approach, one may alternatively wish to re-examine the nominal regression problem $\min_{\beta} g(\mathbf{y} - \mathbf{X}\beta)$ and instead attempt to solve this taking into account noise in the data matrix \mathbf{X} . This approach often takes the form

$$\min_{\beta} \max_{\Delta \in \mathcal{U}} g(\mathbf{y} - (\mathbf{X} + \Delta)\beta),$$

where the set $\mathcal{U} \subseteq \mathbb{R}^{m \times n}$ characterizes the user's belief about uncertainty on the data matrix \mathbf{X} . This set \mathcal{U} is known in the language of robust optimization [2, 5] as an uncertainty set and the inner maximization problem $\max_{\Delta \in \mathcal{U}} g(\mathbf{y} - (\mathbf{X} + \Delta)\beta)$ takes into account the worst-case error (measured via g) over \mathcal{U} . We call such a procedure *robustification* because it attempts to immunize or robustify the regression problem from structural uncertainty in the data. Such a procedure is one of the key tenets of the area of robust optimization [2, 5].

A natural choice of an uncertainty set which gives rise to interpretability is the set $\mathcal{U} = \{\Delta \in \mathbb{R}^{m \times n} : \|\Delta\| \leq \lambda\}$, where $\|\cdot\|$ is some matrix norm and $\lambda > 0$. One can then write $\max_{\Delta \in \mathcal{U}} g(\mathbf{y} - (\mathbf{X} + \Delta)\beta)$ as

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \quad & g(\mathbf{y} - \tilde{\mathbf{X}}\beta) \\ \text{s. t.} \quad & \|\mathbf{X} - \tilde{\mathbf{X}}\| \leq \lambda, \end{aligned}$$

or the worst case error taken over all $\tilde{\mathbf{X}}$ sufficiently close to the data matrix \mathbf{X} . In what follows, if $\|\cdot\|$ is a norm or seminorm, then we let $\mathcal{U}_{\|\cdot\|}$ denote the ball of radius λ in $\|\cdot\|$:

$$\mathcal{U}_{\|\cdot\|} = \{\Delta : \|\Delta\| \leq \lambda\}.$$

For example, \mathcal{U}_{F_p} , \mathcal{U}_{σ_p} , and $\mathcal{U}_{(h,g)}$ denote uncertainty sets under the norms F_p , σ_p , and (h, g) , respectively. We assume $\lambda > 0$ fixed for the remainder of the paper.

2.3 Equivalence of robustification and regularization

A natural question is when do the procedures of regularization and robustification coincide. This problem was first studied in [21] in the context of uncertain least squares problems and has been

extended to more general settings in [31] and most comprehensively in [2]. In this subsection, we present settings in which robustification is equivalent to regularization.

We begin with a general result on robustification under induced seminorm uncertainty sets.

Theorem 1. If $g : \mathbb{R}^m \rightarrow \mathbb{R}$ is a seminorm which is not identically zero and $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a norm, then for any $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(h,g)}} g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta}),$$

where $\mathcal{U}_{(h,g)} = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_{(h,g)} \leq \lambda\}$.

Proof. From the triangle inequality $g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) \leq g(\mathbf{z}) + g(\boldsymbol{\Delta}\boldsymbol{\beta}) \leq g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$ for any $\boldsymbol{\Delta} \in \mathcal{U} := \mathcal{U}_{(h,g)}$. We next show that there exists some $\boldsymbol{\Delta} \in \mathcal{U}$ so that $g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$. Let $\mathbf{v} \in \mathbb{R}^n$ so that $\mathbf{v} \in \operatorname{argmax}_{h^*(\mathbf{v})=1} \mathbf{v}'\boldsymbol{\beta}$, where h^* is the dual norm of h . Note in particular that $\mathbf{v}'\boldsymbol{\beta} = h(\boldsymbol{\beta})$ by the definition of the dual norm h^* . For now suppose that $g(\mathbf{z}) \neq 0$. Define the rank one matrix $\hat{\boldsymbol{\Delta}} = \frac{\lambda}{g(\mathbf{z})} \mathbf{z}\mathbf{v}'$. Observe that

$$g(\mathbf{z} + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = g\left(\mathbf{z} + \frac{\lambda h(\boldsymbol{\beta})}{g(\mathbf{z})} \mathbf{z}\right) = \frac{g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})}{g(\mathbf{z})} g(\mathbf{z}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta}).$$

We next show that $\hat{\boldsymbol{\Delta}} \in \mathcal{U}$. Observe that for any $\mathbf{x} \in \mathbb{R}^n$ that

$$g(\hat{\boldsymbol{\Delta}}\mathbf{x}) = g\left(\frac{\lambda \mathbf{v}'\mathbf{x}}{g(\mathbf{z})} \mathbf{z}\right) = \lambda |\mathbf{v}'\mathbf{x}| \leq \lambda h(\mathbf{x}) h^*(\mathbf{v}) = \lambda h(\mathbf{x}),$$

where the final inequality follows by definition of the dual norm. Hence $\hat{\boldsymbol{\Delta}} \in \mathcal{U}$, as desired.

We now consider the case when $g(\mathbf{z}) = 0$. Let $\mathbf{u} \in \mathbb{R}^m$ so that $g(\mathbf{u}) = 1$ (because g is not identically zero there exists some \mathbf{u} so that $g(\mathbf{u}) > 0$, and so by homogeneity of g we can take \mathbf{u} so that $g(\mathbf{u}) = 1$). Let \mathbf{v} be as before. Now define $\hat{\boldsymbol{\Delta}} = \lambda \mathbf{u}\mathbf{v}'$. We observe that

$$g(\mathbf{z} + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = g(\mathbf{z} + \lambda \mathbf{u}\mathbf{v}'\boldsymbol{\beta}) \leq g(\mathbf{z}) + \lambda |\mathbf{v}'\boldsymbol{\beta}| g(\mathbf{u}) = \lambda h(\boldsymbol{\beta}).$$

Now, by the reverse triangle inequality,

$$g(\mathbf{z} + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) \geq g(\hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) - g(\mathbf{z}) = g(\hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = \lambda h(\boldsymbol{\beta}),$$

and therefore $g(\mathbf{z} + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = \lambda h(\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$. The proof that $\hat{\boldsymbol{\Delta}} \in \mathcal{U}$ is identical to the case when $g(\mathbf{z}) \neq 0$. This completes the proof. \square

This result implies as a corollary known results on the connection between robustification and regularization as found in [31, 2].

Corollary 1. If $p, q \in [1, \infty]$ then

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,p)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda \|\boldsymbol{\beta}\|_q.$$

In particular, for $p = q = 2$ we recover regularized least squares as a robustification; likewise, for $p = 2$ and $q = 1$ we recover Lasso.

In certain special other cases it is also known that robustification is equivalent to regularization. We summarize these in the following theorem. While these results do not follow from Theorem 1, the proofs are similar.

Theorem 2 ([31, 2]). One has the following for any $p, q \in [1, \infty]$:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_p}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda \|\boldsymbol{\beta}\|_{p^*},$$

where p^* is the conjugate of p . Similarly,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_2.$$

Observe that regularized least squares arises again under all uncertainty sets defined by the spectral norms σ_q when the loss function is $g = \ell_2$. Now we continue with a remark on how Lasso arises through regularization.

Remark 1. As per Corollary 1 it is known that Lasso arises as uncertain ℓ_2 regression with uncertainty set $\mathcal{U} := \mathcal{U}_{(1,2)}$. As with Theorem 1, one might argue that the ℓ_1 penalizer arises as an artifact of the model of uncertainty. We remark that one can derive the set \mathcal{U} as an induced uncertainty set defined using the “true” non-convex penalty ℓ_0 , where $\|\boldsymbol{\beta}\|_0 := |\{i : \beta_i \neq 0\}|$. To be precise, for any $p \in [1, \infty]$ and for $\Gamma = \{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_p \leq 1\}$ we claim that

$$\mathcal{U}' := \left\{ \boldsymbol{\Delta} : \max_{\boldsymbol{\beta} \in \Gamma} \frac{\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_0} \leq \lambda \right\}$$

satisfies $\mathcal{U} = \mathcal{U}'$. This is summarized, with an additional representation \mathcal{U}'' as used in [31], in the following proposition.

Proposition 1. If $\mathcal{U} = \mathcal{U}_{(1,2)}$, $\mathcal{U}' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \leq \lambda \|\boldsymbol{\beta}\|_0 \ \forall \|\boldsymbol{\beta}\|_p \leq 1\}$ and $\mathcal{U}'' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}_i\|_2 \leq \lambda \ \forall i\}$, where $\boldsymbol{\Delta}_i$ is the i th column of $\boldsymbol{\Delta}$, then $\mathcal{U} = \mathcal{U}' = \mathcal{U}''$.

Proof. We first show that $\mathcal{U} = \mathcal{U}'$. Because $\|\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}\|_0$ for all $\boldsymbol{\beta} \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}\|_p \leq 1$, we have that $\mathcal{U} \subseteq \mathcal{U}'$. Now suppose that $\boldsymbol{\Delta} \in \mathcal{U}'$. Then for any $\boldsymbol{\beta} \in \mathbb{R}^n$, we have that

$$\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 = \left\| \sum_i \beta_i \boldsymbol{\Delta} \mathbf{e}_i \right\|_2 \leq \sum_i |\beta_i| \|\boldsymbol{\Delta} \mathbf{e}_i\|_2 \leq \sum_i |\beta_i| \lambda = \lambda \|\boldsymbol{\beta}\|_1,$$

where $\{\mathbf{e}_i\}_{i=1}^n$ is the standard orthonormal basis for \mathbb{R}^n . Hence, $\boldsymbol{\Delta} \in \mathcal{U}$ and therefore $\mathcal{U}' \subseteq \mathcal{U}$. Combining with the previous direction gives $\mathcal{U} = \mathcal{U}'$.

We now prove that $\mathcal{U} = \mathcal{U}''$. That $\mathcal{U}'' \subseteq \mathcal{U}$ is essentially obvious; $\mathcal{U} \subseteq \mathcal{U}''$ follows by considering $\boldsymbol{\beta} \in \{\mathbf{e}_i\}_{i=1}^n$. This completes the proof. \square

This proposition implies that ℓ_1 arises from the robustification setting without appealing to standard convexity arguments for why ℓ_1 should be used to replace ℓ_0 (which use the fact that ℓ_1 is the so-called convex envelope of ℓ_0 on $[-1, 1]^n$, see e.g. [11]).

We continue this subsection with another example of when robustification is equivalent to regularization for the case of LAD (ℓ_1) and maximum absolute deviation (ℓ_∞) regression under row-wise uncertainty.

Theorem 3. Fix $q \in [1, \infty]$ and let $\mathcal{U} = \{\boldsymbol{\Delta} : \|\boldsymbol{\delta}_i\|_q \leq \lambda \ \forall i\}$, where $\boldsymbol{\delta}_i$ is the i th row of $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$. Then

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_1 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 + m\lambda \|\boldsymbol{\beta}\|_{q^*}$$

and

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_\infty = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_\infty + \lambda \|\boldsymbol{\beta}\|_{q^*}.$$

Proof. For any $\Delta \in \mathcal{U}$, observe that $\|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_1 \leq \|\mathbf{y} - \mathbf{X}\beta\|_1 + \|\Delta\beta\|_1$. Now,

$$\begin{aligned} \|\Delta\beta\|_1 &= \sum_i |\delta'_i \beta| \\ &\leq \sum_i \|\delta_i\|_q \|\beta\|_{q^*} \quad \text{by Hölder's inequality} \\ &= \|\beta\|_{q^*} \sum_i \|\delta_i\|_q \\ &\leq \|\beta\|_{q^*} \sum_i \lambda \\ &= m\lambda \|\beta\|_{q^*}. \end{aligned}$$

Note that taking any $\delta \in \operatorname{argmax}_{\mathbf{x}: \|\mathbf{x}\|_q=1} \mathbf{x}'\beta$ and setting $\delta_i = \operatorname{sign}(\mathbf{x}'_i \beta - y_i) \lambda \delta$ for all i makes all of the inequalities into equalities (here \mathbf{x}_i is the i th row of \mathbf{X}). Hence,

$$\max_{\Delta \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_1 = \|\mathbf{y} - \mathbf{X}\beta\|_1 + m\lambda \|\beta\|_{q^*}.$$

The proof where ℓ_1 is replaced by ℓ_∞ is similar. \square

For completeness, we note that the uncertainty set $\mathcal{U} = \{\Delta : \|\delta_i\|_q \leq \lambda \forall i\}$ considered in Theorem 3 is actually an induced uncertainty set, namely,

$$\mathcal{U} = \mathcal{U}_{(q^*, \infty)}.$$

In the final part of this subsection we use these results to comment on dictionary learning.

Dictionary learning and robustification

Dictionary learning is concerned with sparse representations of a given collection of vectors $\{\mathbf{x}_i\}_{i=1}^m \subseteq \mathbb{R}^n$ in an unknown, possibly overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{n \times \ell}$. Dictionary learning problems are relevant in a variety of domains, including image processing and genomics [25, 24]. The underlying non-convex problem typically takes a form such as

$$\begin{aligned} \min_{\mathbf{D}, \{\alpha_i\}_{i=1}^m} \quad & \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2 \\ \text{s. t.} \quad & \|\alpha_i\|_0 \leq k \quad \forall i \\ & \|\mathbf{D}_j\|_2 \leq 1 \quad \forall j, \end{aligned}$$

with the constraint $\|\mathbf{D}_j\|_2 \leq 1$ on the column norms of \mathbf{D} controlling the complexity of the dictionary. This is essentially in direct analogy with the usual least squares problem with an additional ℓ_0 constraint, except now \mathbf{D} is unknown (in contrast to before where \mathbf{X} was a known data matrix).

As with other problems the non-convex ℓ_0 constraints are often convexified and placed in the objective to yield

$$\begin{aligned} \min_{\mathbf{D}, \{\alpha_i\}_{i=1}^m} \quad & \sum_{i=1}^m (\|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2 + \lambda \|\alpha_i\|_1) \\ \text{s. t.} \quad & \|\mathbf{D}_j\|_2 \leq 1 \quad \forall j, \end{aligned} \tag{1}$$

which is still a non-convex problem, but one for which large-scale heuristic techniques are known (observe that the problem is convex in \mathbf{D} given fixed $\{\alpha_i\}_{i=1}^m$, and convex in $\{\alpha_i\}_{i=1}^m$ given fixed \mathbf{D} ; it is therefore nicely amenable to techniques such as alternating minimization, see e.g. [24]).

Given the preceding results on vector regression it is not surprising that the dictionary learning problem (1) can be written as an uncertain version of the problem

$$\begin{aligned} \min_{\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^m} \quad & \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \\ \text{s. t.} \quad & \|\mathbf{D}_j\|_2 \leq 1 \quad \forall j, \end{aligned}$$

subject to uncertainty in the unknown matrix \mathbf{D} . Before providing an interpretation, we first set up the uncertainty sets. We assume that each \mathbf{x}_i is composed according to “atoms” in the dictionary \mathbf{D} , with (possibly) different uncertainty in \mathbf{D} for each \mathbf{x}_i :

$$\mathbf{x}_i \approx (\mathbf{D} + \boldsymbol{\Delta}^{(i)})\boldsymbol{\alpha}_i,$$

where $\boldsymbol{\Delta}^{(i)} \in \mathcal{U}^{(i)} \subseteq \mathbb{R}^{n \times \ell}$. Then the uncertain problem appears as

$$\begin{aligned} \min_{\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^m} \quad & \sum_{i=1}^m \left(\max_{\boldsymbol{\Delta}^{(i)} \in \mathcal{U}^{(i)}} \|\mathbf{x}_i - (\mathbf{D} + \boldsymbol{\Delta}^{(i)})\boldsymbol{\alpha}_i\|_2 \right) \\ \text{s. t.} \quad & \|\mathbf{D}_j\|_2 \leq 1 \quad \forall j. \end{aligned} \tag{2}$$

As per Corollary 1, we arrive at the following result for dictionary learning:

Corollary 2. The dictionary learning problem (1) can be written as the uncertain problem (2) with uncertainty set choice $\mathcal{U}^{(i)} = \mathcal{U}_{(1,2)}$ for all $i = 1, \dots, m$.

This gives an alternative interpretation of dictionary learning. Namely, given some fixed dictionary \mathbf{D} , the signals \mathbf{x}_i are composed according to \mathbf{D} with noise in the underlying dictionary occurring independently across signals. As with Lasso, this interpretation is in contrast to the usual sparsity-driven explanations for why dictionary learning arises in the form (1).

2.4 Non-equivalence of robustification and regularization

In contrast to previous work studying robustification for regression, which primarily addresses tractability of solving the new uncertain problem [2] or the implications for Lasso [31], we instead focus our attention on characterization of the equivalence between robustification and regularization. We begin with a regularization upper bound on robustification problems.

Proposition 2. Let $\mathcal{U} \subseteq \mathbb{R}^{m \times n}$ be any non-empty, compact set and $g : \mathbb{R}^m \rightarrow \mathbb{R}$ a seminorm. Then there exists some seminorm $\bar{h} : \mathbb{R}^n \rightarrow \mathbb{R}$ so that for any $\mathbf{z} \in \mathbb{R}^m$, $\boldsymbol{\beta} \in \mathbb{R}^n$,

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) \leq g(\mathbf{z}) + \bar{h}(\boldsymbol{\beta}),$$

with equality when $\mathbf{z} = \mathbf{0}$.

Proof. Let $\bar{h} : \mathbb{R}^n \rightarrow \mathbb{R}$ be defined as

$$\bar{h}(\boldsymbol{\beta}) := \max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\beta}).$$

To show that \bar{h} is a seminorm we must show it satisfies absolute homogeneity and the triangle inequality. For any $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$\bar{h}(\alpha\boldsymbol{\beta}) = \max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\boldsymbol{\Delta}(\alpha\boldsymbol{\beta})) = \max_{\boldsymbol{\Delta} \in \mathcal{U}} |\alpha| g(\boldsymbol{\Delta}\boldsymbol{\beta}) = |\alpha| \left(\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\beta}) \right) = |\alpha| \bar{h}(\boldsymbol{\beta}),$$

so absolute homogeneity is satisfied. Similarly, if $\beta, \gamma \in \mathbb{R}^n$,

$$\bar{h}(\beta + \gamma) = \max_{\Delta \in \mathcal{U}} g(\Delta(\beta + \gamma)) \leq \max_{\Delta \in \mathcal{U}} [g(\Delta\beta) + g(\Delta\gamma)] \leq \left(\max_{\Delta \in \mathcal{U}} g(\Delta\beta) \right) + \left(\max_{\Delta \in \mathcal{U}} g(\Delta\gamma) \right),$$

and hence the triangle inequality is satisfied. Therefore, \bar{h} is a seminorm which satisfies the desired properties, completing the proof. \square

When equality is attained for all pairs $(\mathbf{z}, \beta) \in \mathbb{R}^m \times \mathbb{R}^n$, we are in the regime of the previous subsection, and we say that robustification under \mathcal{U} is equivalent to regularization under \bar{h} (of course, \bar{h} is the only possible candidate for such an equivalency). We now discuss a variety of explicit settings in which regularization only provides upper and lower bounds to the true robustified problem.

Fix $p, q \in [1, \infty]$. Consider the robust ℓ_p regression problem

$$\min_{\beta} \max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p,$$

where $\mathcal{U}_{F_q} = \{\Delta \in \mathbb{R}^{m \times n} : \|\Delta\|_{F_q} \leq \lambda\}$. In the case when $p = q$ we saw earlier (Theorem 2) that one exactly recovers ℓ_p regression with an ℓ_{p^*} penalty:

$$\min_{\beta} \max_{\Delta \in \mathcal{U}_{F_p}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_p + \lambda \|\beta\|_{p^*}.$$

Let us now consider the case when $p \neq q$. We claim that regularization (with \bar{h}) is no longer equivalent to robustification (with \mathcal{U}_{F_q}) unless $p \in \{1, \infty\}$. Applying Proposition 2, one has for any $\mathbf{z} \in \mathbb{R}^m$ that

$$\max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{z} + \Delta\beta\|_p \leq \|\mathbf{z}\|_p + \bar{h}(\beta),$$

where $\bar{h} = \max_{\Delta \in \mathcal{U}_{F_q}} \|\Delta\beta\|_p$ is a norm (when $p = q$, this is precisely the ℓ_{p^*} norm). Here we can compute \bar{h} . To do this we first define a discrepancy function as follows:

Definition 1. For $a, b \in [1, \infty]$ define the discrepancy function $\delta_m(a, b)$ as

$$\delta_m(a, b) := \max\{\|\mathbf{u}\|_a : \mathbf{u} \in \mathbb{R}^m, \|\mathbf{u}\|_b = 1\}.$$

This discrepancy function is computable and well-known, see e.g. [23]. It satisfies $1 \leq \delta_m(a, b) \leq m$ and $\delta_m(a, b)$ is continuous in a and b . One has that $\delta_m(a, b) = \delta_m(b, a) = 1$ if and only if $a = b$ (so long as $m \geq 2$). Using this, we now proceed with the theorem.

Theorem 4. (a) For any $\mathbf{z} \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$,

$$\max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{z} + \Delta\beta\|_p \leq \|\mathbf{z}\|_p + \lambda \delta_m(p, q) \|\beta\|_{q^*}. \quad (3)$$

(b) When $p \in \{1, \infty\}$, there is equality in (3) for all (\mathbf{z}, β) .

(c) When $p \in (1, \infty)$ and $p \neq q$, for any $\beta \neq \mathbf{0}$ the set of $\mathbf{z} \in \mathbb{R}^m$ for which the inequality (3) holds at equality is a finite union of one-dimensional subspaces (so long as $m \geq 2$). Hence, for any $\beta \neq \mathbf{0}$ the inequality in (3) is strict for almost all \mathbf{z} .

(d) For $p \in (1, \infty)$, one has for all $\mathbf{z} \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$ that

$$\|\mathbf{z}\|_p + \frac{\lambda}{\delta_m(q, p)} \|\beta\|_{q^*} \leq \max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{z} + \Delta\beta\|_p. \quad (4)$$

- (e) For $p \in (1, \infty)$, the bound in (4) is best possible in the sense that the gap can be arbitrarily small, i.e., for any $\beta \in \mathbb{R}^n$,

$$\inf_{\mathbf{z}} \left(\max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{z} + \Delta\beta\|_p - \|\mathbf{z}\|_p - \frac{\lambda}{\delta_m(q, p)} \|\beta\|_{q^*} \right) = 0.$$

Proof. (a) We begin by proving the upper bound. Here we proceed by showing that \bar{h} above is precisely $\bar{h}(\beta) = \lambda\delta_m(p, q)\|\beta\|_{q^*}$. Now observe that for any $\Delta \in \mathcal{U}_{F_q}$,

$$\|\Delta\beta\|_p \leq \delta_m(p, q)\|\Delta\beta\|_q \leq \delta_m(p, q)\|\Delta\|_{F_q}\|\beta\|_{q^*} \leq \delta_m(p, q)\lambda\|\beta\|_{q^*}. \quad (5)$$

The first inequality follows by the definition of the discrepancy function δ_m . The second inequality follows from a well-known matrix inequality: $\|\Delta\beta\|_q \leq \|\Delta\|_{F_q}\|\beta\|_{q^*}$ (this follows from a simple application of Hölder's inequality). Now observe that in the chain of inequalities in (5), if one takes any $\mathbf{u} \in \operatorname{argmax} \delta_m(p, q)$ and any $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1} \mathbf{v}'\beta$, then $\hat{\Delta} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and $\|\hat{\Delta}\beta\|_p = \delta_m(p, q)\lambda\|\beta\|_{q^*}$. Hence, $\bar{h}(\beta) = \delta_m(p, q)\lambda\|\beta\|_{q^*}$. This proves the upper bound.

- (b) We now prove that for $p \in \{1, \infty\}$ that one has equality for all $(\mathbf{z}, \beta) \in \mathbb{R}^m \times \mathbb{R}^n$. This follows an argument similar to that given in Theorem 1. First consider the case when $p = 1$. Fix $\mathbf{z} \in \mathbb{R}^m$. Again let $\mathbf{u} \in \operatorname{argmax} \delta_m(1, q)$ and $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1} \mathbf{v}'\beta$. Without loss of generality we may assume that $\operatorname{sign}(z_i) = \operatorname{sign}(u_i)$ for $i = 1, \dots, m$ (one may change the sign of entries of \mathbf{u} and it is still in $\operatorname{argmax} \delta_m(1, q)$). Then again we have $\hat{\Delta} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and

$$\|\mathbf{z} + \hat{\Delta}\beta\|_1 = \|\mathbf{z} + \lambda\mathbf{u}\mathbf{v}'\beta\|_1 = \|\mathbf{z} + \lambda\|\beta\|_{q^*}\mathbf{u}\|_1 = \|\mathbf{z}\|_1 + \lambda\|\beta\|_{q^*}\|\mathbf{u}\|_1 = \|\mathbf{z}\|_1 + \lambda\|\beta\|_{q^*}\delta_m(1, q).$$

Hence, one has equality in the upper bound for $p = 1$, as claimed.

We now turn our attention to the case $p = \infty$. Note that $\delta_m(\infty, q) = 1$ because $\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_q$ for all $\mathbf{z} \in \mathbb{R}^m$. Fix $\mathbf{z} \in \mathbb{R}^m$, and again let $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1} \mathbf{v}'\beta$. Let $\ell \in \{1, \dots, m\}$ so that $|z_\ell| = \|\mathbf{z}\|_\infty$. Define $\mathbf{u} = \operatorname{sign}(z_\ell)\mathbf{e}_\ell \in \mathbb{R}^m$, where \mathbf{e}_ℓ is the vector whose only nonzero entry is a 1 in the ℓ th position. Now observe that $\hat{\Delta} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and

$$\|\mathbf{z} + \hat{\Delta}\beta\|_\infty = \|\mathbf{z} + \operatorname{sign}(z_\ell)\lambda\|\beta\|_{q^*}\mathbf{e}_\ell\|_\infty = \|\mathbf{z}\|_\infty + \lambda\|\beta\|_{q^*}\|\mathbf{e}_\ell\|_\infty = \|\mathbf{z}\|_\infty + \lambda\|\beta\|_{q^*},$$

which proves equality in (3), as was to be shown.

- (c) To proceed, we examine the case where $p \in (1, \infty)$ and consider for which (\mathbf{z}, β) the inequality in (3) is strict. Fix $\beta \neq \mathbf{0}$. For $p \in (1, \infty)$ and $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m$, one has by Minkowski's inequality that $\|\mathbf{y} + \mathbf{z}\|_p = \|\mathbf{y}\|_p + \|\mathbf{z}\|_p$ if and only if one of \mathbf{y} or \mathbf{z} is a non-negative scalar multiple of the other. To have equality in (3), it must be that there exists some $\Delta \in \operatorname{argmax}_{\Delta \in \mathcal{U}_{F_q}} \|\Delta\beta\|_p$ for which $\|\mathbf{z} + \Delta\beta\|_p = \|\mathbf{z}\|_p + \|\Delta\beta\|_p$. For any $\mathbf{z} \neq \mathbf{0}$ this observation, combined with Minkowski's inequality, implies that

$$\|\Delta\|_{F_q} = \lambda, \quad \Delta\beta = \mu\mathbf{z} \quad \text{for some } \mu \geq 0, \quad \text{and} \quad \|\Delta\beta\|_p = \lambda\delta_m(p, q)\|\beta\|_{q^*}.$$

The first and last equalities imply that $\Delta\beta \in \lambda\|\beta\|_{q^*} \operatorname{argmax} \delta_m(p, q)$. Note that $\operatorname{argmax} \delta_m(p, q)$ is finite whenever $p \neq q$ and $m \geq 2$, a geometric property of ℓ_p balls. Hence, taking any \mathbf{z} which is not a scalar multiple of a point in $\operatorname{argmax} \delta_m(p, q)$ implies by Minkowski's inequality that

$$\max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{z} + \Delta\beta\|_p < \|\mathbf{z}\|_p + \lambda\delta_m(p, q)\|\beta\|_{q^*}.$$

Hence, for any $\beta \neq \mathbf{0}$, the inequality in (3) is strict for all \mathbf{z} not in a finite union of one-dimensional subspaces, so long as $p \in (1, \infty)$, $p \neq q$, and $m \geq 2$.

- (d) We now prove the lower bound in (4). We first prove that this is a valid lower bound. Let $\mathbf{v} \in \mathbb{R}^n$ so that

$$\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1} \mathbf{v}'\boldsymbol{\beta}.$$

Hence $\mathbf{v}'\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_{q^*}$ by the definition of the dual norm. Define $\hat{\boldsymbol{\Delta}} = \frac{\lambda}{\|\mathbf{z}\|_q} \mathbf{z}\mathbf{v}'$. Observe that $\hat{\boldsymbol{\Delta}} \in \mathcal{U}_{F_q}$. Further, note that $\|\mathbf{z}\|_q \leq \delta_m(q, p)\|\mathbf{z}\|_p$ by definition of δ_m and therefore $1/\delta_m(q, p) \leq \|\mathbf{z}\|_p/\|\mathbf{z}\|_q$. Putting things together,

$$\|\mathbf{z}\|_p + \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}}{\delta_m(q, p)} \leq \|\mathbf{z}\|_p + \frac{\lambda\|\mathbf{z}\|_p\|\boldsymbol{\beta}\|_{q^*}}{\|\mathbf{z}\|_q} = \|\mathbf{z}\|_p \left(1 + \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}}{\|\mathbf{z}\|_p}\right) = \|\mathbf{z} + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}\|_p \leq \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p.$$

This completes the proof of the lower bound.

- (e) To conclude we prove that the gap in (4) can be made arbitrarily small for $p \in (1, \infty)$. We proceed in several steps. We first prove that for any $\mathbf{z} \neq \mathbf{0}$ that

$$\lim_{\alpha \rightarrow \infty} \left(\max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_q}} \|\alpha\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p - \|\alpha\mathbf{z}\|_p \right) = \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}}, \quad (6)$$

where we use the shorthand \mathbf{z}^{p-1} to denote the vector in \mathbb{R}^m whose i th entry is $|z_i|^{p-1}$. Observe that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_q}} \|\alpha\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \|\alpha\mathbf{z} + \mathbf{u}\|_p$$

(see Proposition 5). It is easy to argue that $\mathbf{u} \in \operatorname{argmax}_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \|\alpha\mathbf{z} + \mathbf{u}\|_p$ has $\operatorname{sign}(u_i) = \operatorname{sign}(\alpha z_i)$ for all i . Therefore, we restrict our attention without loss of generality to $\mathbf{z} \geq \mathbf{0}$, $\mathbf{z} \neq \mathbf{0}$, and $\mathbf{u} \geq \mathbf{0}$. For any \mathbf{u} such that $\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}$ and $\mathbf{u} \geq \mathbf{0}$, note that

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p &= \lim_{\alpha \rightarrow \infty} \frac{\|\mathbf{z} + \mathbf{u}/\alpha\|_p - \|\mathbf{z}\|_p}{1/\alpha} \\ &= \lim_{\bar{\alpha} \rightarrow 0^+} \frac{\|\mathbf{z} + \bar{\alpha}\mathbf{u}\|_p - \|\mathbf{z}\|_p}{\bar{\alpha}} \\ &= \frac{d}{d\bar{\alpha}} \bigg|_{\bar{\alpha}=0} \|\mathbf{z} + \bar{\alpha}\mathbf{u}\|_p \\ &= \frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}}. \end{aligned}$$

We can now proceed to finish the claim in (6) (still restricting attention to $\mathbf{z} \geq \mathbf{0}$ without loss of generality). By the above arguments, for any $\mathbf{u} \geq \mathbf{0}$ and any $\epsilon > 0$ there exists some $\hat{\alpha} = \hat{\alpha}(\mathbf{u}) > 0$ sufficiently large so that for all $\alpha > \hat{\alpha}$,

$$\left| \|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p - \frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}} \right| \leq \epsilon.$$

It remains to be shown that for any $\epsilon > 0$ there exists some $\hat{\alpha}$ so that for all $\alpha > \hat{\alpha}$,

$$\left| \left(\max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p \right) - \left(\max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}} \right) \right| \leq \epsilon.$$

We prove this as follows. Let $\epsilon > 0$. Choose points $\{\mathbf{u}_1, \dots, \mathbf{u}_M\} \subseteq \mathbb{R}^m$ with $\|\mathbf{u}_j\|_q = \lambda \|\boldsymbol{\beta}\|_{q^*} \forall j$ so that for any $\mathbf{u} \in \mathbb{R}^m$ with $\|\mathbf{u}\|_q = \lambda \|\boldsymbol{\beta}\|_{q^*}$, there exists some j so that $\|\mathbf{u} - \mathbf{u}_j\|_p \leq \epsilon/3$. Now observe that for any α ,

$$\begin{aligned} \max_j \|\alpha \mathbf{z} + \mathbf{u}_j\|_p &\leq \max_{\|\mathbf{u}\|_q \leq \lambda \|\boldsymbol{\beta}\|_{q^*}} \|\alpha \mathbf{z} + \mathbf{u}\|_p \\ &\leq \max_j \left(\max_{\|\mathbf{u} - \mathbf{u}_j\|_p \leq \epsilon/3} \|\alpha \mathbf{z} + \mathbf{u}\|_p \right) \\ &= \max_j \left(\max_{\|\bar{\mathbf{u}}\|_p \leq \epsilon/3} \|\alpha \mathbf{z} + \mathbf{u}_j + \bar{\mathbf{u}}\|_p \right) \\ &\leq \max_j \left(\max_{\|\bar{\mathbf{u}}\|_p \leq \epsilon/3} \|\alpha \mathbf{z} + \mathbf{u}_j\|_p + \|\bar{\mathbf{u}}\|_p \right) \\ &= \epsilon/3 + \max_j \|\alpha \mathbf{z} + \mathbf{u}_j\|_p. \end{aligned}$$

Similarly, one has for $\bar{\mathbf{z}} = \mathbf{z}^{p-1}/\|\mathbf{z}\|_p^{p-1}$ that $\left| \max_j \mathbf{u}'_j \bar{\mathbf{z}} - \max_{\|\mathbf{u}\|_q \leq \lambda \|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}' \bar{\mathbf{z}} \right| \leq \epsilon/3$. Now for each j choose $\hat{\alpha}_j$ so that for all $\alpha > \hat{\alpha}_j$, $\left| \|\alpha \mathbf{z} + \mathbf{u}_j\|_p - \|\alpha \mathbf{z}\|_p - \mathbf{u}'_j \mathbf{z}^{p-1}/\|\mathbf{z}\|_p^{p-1} \right| \leq \epsilon/3$. Define $\hat{\alpha} = \max_j \hat{\alpha}_j$. Now observe that by combining the above two observations, one has for any $\alpha > \hat{\alpha}$ that

$$\begin{aligned} \left| \left(\max_{\|\mathbf{u}_q\| \leq \lambda \|\boldsymbol{\beta}\|_{q^*}} \|\alpha \mathbf{z} + \mathbf{u}\|_p - \|\alpha \mathbf{z}\|_p \right) - \left(\max_{\|\mathbf{u}\|_q \leq \lambda \|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}' \bar{\mathbf{z}} \right) \right| &\leq \\ &\leq 2\epsilon/3 + \left| \left(\max_j \|\alpha \mathbf{z} + \mathbf{u}_j\|_p - \|\alpha \mathbf{z}\|_p \right) - \left(\max_{\ell} \mathbf{u}'_{\ell} \bar{\mathbf{z}} \right) \right| \\ &\leq 2\epsilon/3 + \max_j \left| \|\alpha \mathbf{z} + \mathbf{u}_j\|_p - \|\alpha \mathbf{z}\|_p - \mathbf{u}'_j \bar{\mathbf{z}} \right| \\ &\leq 2\epsilon/3 + \epsilon/3 \\ &= \epsilon. \end{aligned}$$

Noting that $\max_{\|\mathbf{u}\|_q \leq \lambda \|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}' \bar{\mathbf{z}} = \lambda \|\boldsymbol{\beta}\|_{q^*} \|\bar{\mathbf{z}}\|_{q^*}$ concludes the proof of (6).

We now claim that

$$\min_{\mathbf{z}} \frac{\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}} = \frac{1}{\delta_m(q, p)}. \quad (7)$$

First note that

$$\min_{\mathbf{z}} \frac{\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}} = \frac{1}{\max_{\mathbf{z}} \frac{\|\mathbf{z}\|_p^{p-1}}{\|\mathbf{z}^{p-1}\|_{q^*}}}.$$

Now observe that

$$\max_{\mathbf{z}} \frac{\|\mathbf{z}\|_p^{p-1}}{\|\mathbf{z}^{p-1}\|_{q^*}} = \max_{\tilde{\mathbf{z}}} \frac{\|\tilde{\mathbf{z}}\|_{p^*}}{\|\tilde{\mathbf{z}}\|_{q^*}} = \delta_m(p^*, q^*). \quad (8)$$

We prove this as follows: given \mathbf{z} , let $\tilde{\mathbf{z}} = \mathbf{z}^{p-1}$. Then one can show that $\|\tilde{\mathbf{z}}\|_{p^*}/\|\mathbf{z}\|_p^{p-1} = 1$, and so $\|\tilde{\mathbf{z}}\|_{p^*}/\|\tilde{\mathbf{z}}\|_{q^*} = \|\mathbf{z}\|_p^{p-1}/\|\mathbf{z}^{p-1}\|_{q^*}$. The converse is similar, proving (8). Finally, note that $\delta_m(p^*, q^*) = \delta_m(q, p)$, which follows by a simply duality argument, and therefore (7) is proven. To finish the argument, pick any $\mathbf{z} \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}^{p-1}\|_{q^*}/\|\mathbf{z}\|_p^{p-1}$. Per (7), $\|\mathbf{z}^{p-1}\|_{q^*}/\|\mathbf{z}\|_p^{p-1} =$

$1/\delta_m(q, p)$. Hence, now applying (6), given any $\epsilon > 0$, there exists some $\alpha > 0$ large enough so that

$$\left| \left(\max_{\Delta \in \mathcal{U}_{F_q}} \|\alpha \mathbf{z} + \Delta \beta\|_p \right) - \left(\|\alpha \mathbf{z}\|_p + \frac{\lambda}{\delta_m(q, p)} \|\beta\|_{q^*} \right) \right| \leq \epsilon.$$

Therefore, the gap in the lower bound in (4) can be made arbitrarily small for any $\beta \in \mathbb{R}^n$. This concludes the proof. \square

Theorem 4 characterizes precisely when robustification under \mathcal{U}_{F_q} is equivalent to regularization for the case of ℓ_p regression. In particular, when $p \neq q$ and $p \in (1, \infty)$, the two are *not* equivalent, and one only has that

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_p + \frac{\lambda}{\delta_m(q, p)} \|\beta\|_{q^*} \leq \min_{\beta} \max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p \leq \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_p + \lambda \delta_m(p, q) \|\beta\|_{q^*}.$$

Further, we have shown that these upper and lower bounds are the *best possible* (Theorem 4, parts (c) and (e)). While ℓ_p regression with uncertainty set \mathcal{U}_{F_q} for $p \neq q$ and $p \in (1, \infty)$ still has both upper and lower bounds which correspond to regularization (with different regularization parameters $\bar{\lambda} \in \left[\frac{\lambda}{\delta_m(q, p)}, \lambda \delta_m(p, q) \right]$), we emphasize that in this case there is no longer the direct connection between the parameter garnering the magnitude of uncertainty (λ) and the parameter for regularization ($\bar{\lambda}$).

Let us remark that in general, lower bounds on $\max_{\Delta \in \mathcal{U}} g(\mathbf{z} + \Delta \beta)$ will depend on the structure of \mathcal{U} and may not exist (except for the trivial lower bound of $g(\mathbf{z})$) in some scenarios. However, it is easy to show that if \mathcal{U} is compact and full-dimensional (i.e., \mathcal{U} has non-empty interior) then there exists some $\underline{\lambda} \in (0, 1]$ so that

$$\max_{\Delta \in \mathcal{U}} g(\mathbf{z} + \Delta \beta) \geq g(\mathbf{z}) + \underline{\lambda} \bar{h}(\beta).$$

We now proceed to analyze another setting in which robustification is not equivalent to regularization. The setting, in line with Theorem 2, is ℓ_p regression under spectral uncertainty sets \mathcal{U}_{σ_q} . As per Theorem 2, one has that

$$\min_{\beta} \max_{\Delta \in \mathcal{U}_{\sigma_q}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_2 = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2 + \lambda \|\beta\|_2$$

for any $q \in [1, \infty]$. This result on the “universality” of RLS under a variety of uncertainty sets relies on the fact that the ℓ_2 norm underlies spectral decompositions; namely, one can write any matrix \mathbf{X} as $\sum_i \mu_i \mathbf{u}_i \mathbf{v}_i'$, where $\{\mu_i\}_i$ are the singular values of \mathbf{X} , $\{\mathbf{u}_i\}_i$ and $\{\mathbf{v}_i\}_i$ are the left and right singular vectors of \mathbf{X} , respectively, and $\|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1$ for all i .

A natural question is what happens when the loss function ℓ_2 , a modeling choice, is replaced by ℓ_p , where $p \in [1, \infty]$. We claim that for $p \notin \{1, 2, \infty\}$, robustification under \mathcal{U}_{σ_q} is no longer equivalent to regularization. In light of Theorem 4 this is not difficult to prove. We find that the choice of $q \in [1, \infty]$, as before, is inconsequential. We summarize this in the following proposition:

Proposition 3. For any $\mathbf{z} \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$,

$$\max_{\Delta \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \Delta \beta\|_p \leq \|\mathbf{z}\|_p + \lambda \delta_m(p, 2) \|\beta\|_2. \quad (9)$$

In particular, if $p \in \{1, 2, \infty\}$, there is equality in (9) for all $(\mathbf{z}, \boldsymbol{\beta})$. If $p \notin \{1, 2, \infty\}$, then for any $\boldsymbol{\beta} \neq \mathbf{0}$ the inequality in (9) is strict for almost all \mathbf{z} (when $m \geq 2$). Further, for $p \notin \{1, 2, \infty\}$ one has the lower bound

$$\|\mathbf{z}\|_p + \frac{\lambda}{\delta_m(2, p)} \|\boldsymbol{\beta}\|_2 \leq \max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p,$$

whose gap is arbitrarily small for all $\boldsymbol{\beta}$.

Proof. This result is Theorem 4 in disguise. This follows by noting that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_2}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p$$

(see Proposition 5) and directly applying the preceding results. \square

We now consider a third setting for ℓ_p regression, this time subject to uncertainty $\mathcal{U}_{(q,r)}$; this is a generalized version of the problems considered in Theorems 1 and 3. From Theorem 1 we know that if $p = r$, then

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,p)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda \|\boldsymbol{\beta}\|_q.$$

Similarly, as per Theorem 3, when $r = \infty$ and $p \in \{1, \infty\}$,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,\infty)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda \delta_m(p, \infty) \|\boldsymbol{\beta}\|_q.$$

Given these results, it is natural to inquire what happens for more general choices of induced uncertainty set $\mathcal{U}_{(q,r)}$. As before with Theorem 4, we have full characterization on the equivalence of robustification and regularization for ℓ_p regression with uncertainty set $\mathcal{U}_{(q,r)}$:

Proposition 4. For any $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$,

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p \leq \|\mathbf{z}\|_p + \lambda \delta_m(p, r) \|\boldsymbol{\beta}\|_q. \quad (10)$$

In particular, if $p \in \{1, \infty\}$, there is equality in (9) for all $(\mathbf{z}, \boldsymbol{\beta})$. If $p \in (1, \infty)$ and $p \neq r$, then for any $\boldsymbol{\beta} \neq \mathbf{0}$ the inequality in (10) is strict for almost all \mathbf{z} (when $m \geq 2$). Further, for $p \in (1, \infty)$ with $p \neq r$ one has the lower bound

$$\|\mathbf{z}\|_p + \frac{\lambda}{\delta_m(r, p)} \|\boldsymbol{\beta}\|_q \leq \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p,$$

whose gap is arbitrarily small for all $\boldsymbol{\beta}$.

Proof. The proofs follows the argument given in the proof of Theorem 4. Here we simply note that now one uses the fact that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\|\mathbf{u}\|_r \leq \lambda \|\boldsymbol{\beta}\|_q} \|\mathbf{z} + \mathbf{u}\|_p$$

(see Proposition 5). \square

We summarize all of the results on linear regression in Table 2.

Loss function	Uncertainty set $\mathcal{U} \subseteq \mathbb{R}^{m \times n}$	$\bar{h}(\boldsymbol{\beta})$	Equivalence if and only if
seminorm g	$\mathcal{U}_{(h,g)}$ (h norm)	$\lambda h(\boldsymbol{\beta})$	always
ℓ_p	\mathcal{U}_{σ_q}	$\lambda \delta_m(p, 2) \ \boldsymbol{\beta}\ _2$	$p \in \{1, 2, \infty\}$
ℓ_p	\mathcal{U}_{F_q}	$\lambda \delta_m(p, q) \ \boldsymbol{\beta}\ _{q^*}$	$p = q$ or $p \in \{1, \infty\}$
ℓ_p	$\mathcal{U}_{(q,r)}$	$\lambda \delta_m(p, r) \ \boldsymbol{\beta}\ _q$	$p = r$ or $p \in \{1, \infty\}$
ℓ_p	$\{\boldsymbol{\Delta} : \ \boldsymbol{\delta}_i\ _q \leq \lambda \forall i\} = \mathcal{U}_{(q^*, \infty)}$	$\lambda m^{1/p} \ \boldsymbol{\beta}\ _{q^*}$	$p \in \{1, \infty\}$

Table 2: Summary of equivalencies for robustification with uncertainty set \mathcal{U} and regularization with penalty \bar{h} , where \bar{h} is as given in Proposition 2. Here by equivalence we mean that for all $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$, $\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{z} + \boldsymbol{\beta}) = g(\mathbf{z}) + \bar{h}(\boldsymbol{\beta})$, where g is the loss function, i.e., the upper bound \bar{h} is also a lower bound. Here δ_m is as in Theorem 4. Throughout $p, q \in [1, \infty]$ and $m \geq 2$. Here $\boldsymbol{\delta}_i$ denotes the i th row of $\boldsymbol{\Delta}$.

3 On the equivalence of robustification and regularization in median regression

In this section, we apply the ideas from the connections between robustification and regularization in the linear regression case to the *least quantile squares* (LQS) regression problem [28]. Given $\boldsymbol{\beta} \in \mathbb{R}^n$ and observed $\mathbf{y} \in \mathbb{R}^m$, we let $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ denote the vector of residuals. Let $\{r_{(i)}\}_{i=1}^m$ be the sorted residuals with

$$|r_{(1)}| \leq |r_{(2)}| \leq \dots \leq |r_{(m)}|.$$

LQS at the q th order corresponds to solving

$$\min_{\boldsymbol{\beta}} |r_{(q)}|.$$

Quantile regression, despite being NP-hard in general [3], is of particular interest because of its desirable finite sample breakdown point properties [17]. Authors in [8] have shown that by coupling first-order methods with mixed integer optimization it is possible to solve LQS problems on the scale of practical interest to provable optimality in a matter of hours, a substantial advance over previous enumerative approaches to solving LQS.

In what follows we show that using the same techniques as in [8] combined with the ideas of robust regression as presented above that it is possible to write mixed integer optimization (MIO) formulations for robust LQS with the same structure as nominal LQS formulations from [8]. We begin by recalling the formulation for q th order LQS:

$$\begin{aligned}
& \min && \gamma \\
& && r_i^+ + r_i^- - \gamma = \bar{\mu}_i - \mu_i, \quad i = 1, \dots, m \\
& && r_i^+ - r_i^- = y_i - \mathbf{x}_i' \boldsymbol{\beta}, \quad i = 1, \dots, m \\
& && \sum_{i=1}^m z_i = q \\
& \text{s. t.} && \gamma \geq \mu_i, \quad i = 1, \dots, m \\
& && \mu_i, \bar{\mu}_i \geq 0, \quad i = 1, \dots, m \\
& && r_i^+, r_i^- \geq 0, \quad i = 1, \dots, m \\
& && (\bar{\mu}_i, \mu_i) : \text{SOS-1}, \quad i = 1, \dots, m \\
& && (r_i^+, r_i^-) : \text{SOS-1}, \quad i = 1, \dots, m \\
& && (z_i, \mu_i) : \text{SOS-1}, \quad i = 1, \dots, m \\
& && z_i \in \{0, 1\}, \quad i = 1, \dots, m.
\end{aligned}$$

Here \mathbf{x}_i denotes the i th row of \mathbf{X} and $(a, b) : \text{SOS-1}$ denotes the constraint that $ab = 0$, which is a common constraint in MIO formulations. There exists a variety of high-quality commercial software, such as **Gurobi** and **CPLEX**, for solving MIO problems (see [10] for a survey).

We assume throughout the remainder of this section that the uncertainty set $\mathcal{U} \subseteq \mathbb{R}^{m \times n}$ is defined as $\mathcal{U} = \{\Delta : \|\Delta\| \leq \lambda\}$, where $\|\cdot\|$ is a norm whose dual $\|\cdot\|_*$ satisfies the following separability assumption (a modified form of that taken in [2]):

$$\text{there exist norms } \phi, \psi \text{ so that } \|\mathbf{u}\mathbf{v}'\|_* = \phi(\mathbf{u})\psi(\mathbf{v}) \quad \text{for all } \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n. \quad (11)$$

All norms which we consider – F_p , σ_p , and (h, g) – satisfy (11). The corresponding (ϕ, ψ) for each norm are as follows, summarized in Table 3:

1. For the p -Frobenius norm F_p , its dual is F_{p^*} and for any $\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{u}\mathbf{v}'\|_{F_p^*} = \|\mathbf{u}\|_{p^*}\|\mathbf{v}\|_{p^*}$. Hence, for F_p , $\phi = \psi = \ell_{p^*}$.
2. For the (h, g) -induced norm, where $h : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}$ are norms, the dual is (h^*, g^*) and therefore for any \mathbf{u}, \mathbf{v} , $\|\mathbf{u}\mathbf{v}'\|_{(h^*, g^*)} = g^*(\mathbf{u})h(\mathbf{v})$, and so $\phi = g^*$ and $\psi = h$. In particular, for the (q, p) -induced norm, $\phi = \ell_{p^*}$ and $\psi = \ell_q$.
3. For the p -spectral norms σ_p , the dual is σ_{p^*} and again for any \mathbf{u}, \mathbf{v} , $\|\mathbf{u}\mathbf{v}'\|_{\sigma_{p^*}} = \|\mathbf{u}\|_2\|\mathbf{v}\|_2$, so $\phi = \psi = \ell_2$.

Norm	ϕ	ψ
F_p	ℓ_{p^*}	ℓ_{p^*}
(h, g)	g^*	h
(q, p)	ℓ_{p^*}	ℓ_q
σ_p	ℓ_2	ℓ_2

Table 3: Common matrix norms and their corresponding (ϕ, ψ) satisfying (11). Here $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}$ are arbitrary norms.

Such a separability assumption in (11) is particularly useful because it allows one to rewrite the inner maximization problem in a different form. This is summarized in the following proposition; its proof follows basic techniques from convex optimization [11, 27].

Proposition 5. (a) If $\|\cdot\| : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a norm, $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is any function, and $\mathbf{z} \in \mathbb{R}^m, \beta \in \mathbb{R}^n$, then

$$\sup_{\|\Delta\| \leq \lambda} f(\mathbf{z} + \Delta\beta) = \sup_{\substack{\mathbf{u} \in \mathbb{R}^m : \\ \max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*} \leq \lambda}} f(\mathbf{z} + \mathbf{u}),$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

(b) In particular, under assumption (11), one has

$$\sup_{\|\Delta\| \leq \lambda} f(\mathbf{z} + \Delta\beta) = \sup_{\phi^*(\mathbf{u}) \leq \lambda\psi(\beta)} f(\mathbf{z} + \mathbf{u}),$$

where ϕ^* is the dual norm of ϕ .

Proof. (a) We prove this in two parts. We begin by showing that

$$\sup_{\|\Delta\| \leq \lambda} f(\mathbf{z} + \Delta\beta) = \sup_{\substack{\mathbf{u} \in \mathbb{R}^m: \\ \min_{\Delta} \{\|\Delta\| : \Delta\beta = \mathbf{u}\} \leq \lambda}} f(\mathbf{z} + \mathbf{u}).$$

If $\|\Delta\| \leq \lambda$ then taking $\mathbf{u} = \Delta\beta$ satisfies the desired properties:

$$\min_{\hat{\Delta}} \{\|\hat{\Delta}\| : \hat{\Delta}\beta = \mathbf{u}\} \leq \|\Delta\| \leq \lambda.$$

Likewise, if \mathbf{u} has $\min_{\Delta} \{\|\Delta\| : \Delta\beta = \mathbf{u}\}$ then taking any Δ with $\|\Delta\| \leq \lambda$ and $\Delta\beta = \mathbf{u}$ satisfies the desired properties. Hence, the first part of the proof is completed.

It remains to be shown that $\min_{\Delta} \{\|\Delta\| : \Delta\beta = \mathbf{u}\} = \max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*}$. We first rewrite $\min_{\Delta} \{\|\Delta\| : \Delta\beta = \mathbf{u}\}$ exactly as

$$\min_{\Delta} \max_{\mu} \|\Delta\| + \mu'(\mathbf{u} - \Delta\beta).$$

It is easy to see that strong duality holds for this problem and therefore

$$\min_{\Delta} \max_{\mu} \|\Delta\| + \mu'(\mathbf{u} - \Delta\beta) = \max_{\mu} \min_{\Delta} \|\Delta\| + \mu'(\mathbf{u} - \Delta\beta).$$

Now observe that $\mu' \Delta\beta = \langle \Delta, \mu\beta' \rangle$ and so $\min_{\Delta} \|\Delta\| - \mu' \Delta\beta = -\max_{\Delta} \langle \Delta, \mu\beta' \rangle - \|\Delta\|$ which we recognize as the convex (Fenchel) conjugate for $\|\cdot\|$ [11]. This problem has a simple solution:

$$\max_{\Delta} \langle \Delta, \mu\beta' \rangle - \|\Delta\| = \begin{cases} 0, & \|\mu\beta'\|_* \leq 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Hence,

$$\max_{\mu} \min_{\Delta} \|\Delta\| + \mu'(\mathbf{u} - \Delta\beta) = \max_{\substack{\mu: \\ \|\mu\beta'\|_* \leq 1}} \mu' \mathbf{u} = \max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*}.$$

(b) In particular, under the assumption (11), it is easy to see that

$$\max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*} = \max_{\mu} \frac{\mu' \mathbf{u}}{\phi(\mu)\psi(\beta)} = \frac{1}{\psi(\beta)} \max_{\mu} \frac{\mu' \mathbf{u}}{\phi(\mu)} = \frac{\phi^*(\mathbf{u})}{\psi(\beta)}.$$

Therefore, $\max_{\mu} \mu' \mathbf{u} / \|\mu\beta'\|_* \leq \lambda$ if and only if $\phi^*(\mathbf{u}) \leq \lambda\psi(\beta)$. This concludes the proof. \square

Before proceeding further, we describe the implications of Proposition 5 for several matrix norms:

1. For the p -Frobenius norms F_p , we have

$$\max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*} = \max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\|_{p^*} \|\beta\|_{p^*}} = \frac{\|\mathbf{u}\|_p}{\|\beta\|_{p^*}}.$$

2. For the (h, g) -induced norms with $\|\cdot\| = \|\cdot\|_{(h, g)}$, we have

$$\max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*} = \max_{\mu} \frac{\mu' \mathbf{u}}{g^*(\mu)h(\beta)} = \frac{g(\mathbf{u})}{h(\beta)}.$$

In particular, for the (q, p) -induced norms,

$$\max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\beta'\|_*} = \frac{\|\mathbf{u}\|_p}{\|\beta\|_q}.$$

3. For the p -spectral norm σ_p with $\|\cdot\| = \sigma_p$, we have that $\|\cdot\|_* = \sigma_{p^*}$ and so

$$\max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu \beta'\|_*} = \max_{\mu} \frac{\mu' \mathbf{u}}{\|\mu\|_2 \|\beta\|_2} = \frac{\|\mathbf{u}\|_2}{\|\beta\|_2}.$$

Using Proposition 5 we now write robust analogues of LQS for different choices of uncertainty sets. We operate under assumption (11) and provide full analyses for two specific cases: $\phi = \ell_1$ and $\phi = \ell_\infty$.

3.1 Uncertainty sets with $\phi = \ell_1$

We begin by stating the main theorem here for different choices of ψ .

Theorem 5. For uncertainty sets $\mathcal{U} = \{\Delta : \|\Delta\| \leq \lambda\}$ defined by a norm which satisfies assumption (11) with $(\phi = \ell_1, \psi)$, robust LQS can be exactly reformulated as the mixed integer optimization problem

$$\begin{aligned} \min \quad & \gamma + \tau \\ & \lambda \psi(\beta) \leq \tau \\ & r_i^+ + r_i^- - \gamma = \bar{\mu}_i - \mu_i, \quad i = 1, \dots, m \\ & r_i^+ - r_i^- = y_i - \mathbf{x}_i' \beta, \quad i = 1, \dots, m \\ & \sum_{i=1}^m z_i = q \\ \text{s. t.} \quad & \gamma \geq \mu_i, \quad i = 1, \dots, m \\ & \mu_i, \bar{\mu}_i \geq 0, \quad i = 1, \dots, m \\ & r_i^+, r_i^- \geq 0, \quad i = 1, \dots, m \\ & (\bar{\mu}_i, \mu_i) : \text{SOS-1}, \quad i = 1, \dots, m \\ & (r_i^+, r_i^-) : \text{SOS-1}, \quad i = 1, \dots, m \\ & (z_i, \mu_i) : \text{SOS-1}, \quad i = 1, \dots, m \\ & z_i \in \{0, 1\}, \quad i = 1, \dots, m. \end{aligned} \tag{12}$$

In particular, for the choice of norms F_∞ (with $\psi = \ell_1$), $(1, \infty)$ ($\psi = \ell_1$), and (∞, ∞) ($\psi = \ell_\infty$), this formulation is a linear mixed integer program. Under norm $(2, \infty)$ ($\psi = \ell_2$), the formulation is a mixed integer second order cone optimization problem (MISOCP) (see e.g. [10]).

Proof. In order to apply Proposition 5 we must compute

$$\max_{\mathbf{u}} \{\text{ord}_q |y_i - \mathbf{x}_i' \beta + u_i| : \|\mathbf{u}\|_\infty \leq \lambda \psi(\beta)\},$$

where ord_q denotes the q th order statistic ($|r_{(q)}|$ in the nominal problem). It is easy to see that

$$\max_{\mathbf{u}} \{\text{ord}_q |y_i - \mathbf{x}_i' \beta + u_i| : \|\mathbf{u}\|_\infty \leq \lambda \psi(\beta)\} = \text{ord}_q \{|y_i - \mathbf{x}_i' \beta| + \lambda \psi(\beta)\} = \lambda \psi(\beta) + |r_{(q)}|.$$

Using the nominal formulation for LQS as given in [8], we arrive at formulation (12). \square

3.1.1 Uncertainty sets with $\phi = \ell_\infty$

The formulation given in the following theorem is structurally different than the nominal formulation from [8], in contrast to formulation (12) as above. As before we state the main theorem:

Theorem 6. For uncertainty sets $\mathcal{U} = \{\Delta : \|\Delta\| \leq \lambda\}$ defined by a norm which satisfies assumption (11) with $(\phi = \ell_\infty, \psi)$, robust LQS can be exactly reformulated as the mixed integer optimization problem

$$\begin{aligned}
& \min_{\nu} && \lambda\psi(\beta) \leq (m - q + 1)\rho - \sum_{i=1}^m \tau_i \\
& \text{s. t.} && \begin{aligned} & \rho - \tau_i \leq \pi_i, & i = 1, \dots, m \\ & \pi_i \geq \nu - a_i, & i = 1, \dots, m \\ & r_i^+ - r_i^- = y_i - \mathbf{x}_i' \beta, & i = 1, \dots, m \\ & a_i = r_i^+ + r_i^-, & i = 1, \dots, m \\ & (\pi_i - \nu + a_i, \pi_i) : \text{SOS-1}, & i = 1, \dots, m \\ & (r_i^+, r_i^-) : \text{SOS-1}, & i = 1, \dots, m \\ & \pi, \tau \geq 0. \end{aligned} \end{aligned} \tag{13}$$

In particular, for uncertainty sets $\mathcal{U} = \{\Delta : \|\Delta\| \leq \lambda\}$ defined under the norms F_1 (with $\psi = \ell_\infty$), $(1, 1)$ ($\psi = \ell_1$), and $(\infty, 1)$ ($\psi = \ell_\infty$), this is a linear mixed integer program. Under norm $(2, 1)$ ($\psi = \ell_2$) this formulation is a MISOCP.

Proof. We proceed as before, but the formulation requires a bit more care. Now we must compute

$$\max_{\mathbf{u}} \{\text{ord}_q |y_i - \mathbf{x}_i' \beta + u_i| : \|\mathbf{u}\|_1 \leq \lambda\psi(\beta)\}.$$

The solution to this problem is a classical answer which involves a so-called waterfilling argument, comparable to that given in information theory for channels with colored Gaussian noise [15]. Given residuals \mathbf{r} the solution to this maximization problem is the unique $\nu \in \mathbb{R}$ such that

$$(\nu - |r_{(q)}|)_+ + (\nu - |r_{(q+1)}|)_+ + \dots + (\nu - |r_{(m)}|)_+ = \lambda\psi(\beta),$$

where $(a)_+ := \max\{0, a\}$; this follows from a direct derivation employing the Karush-Kuhn-Tucker conditions (see e.g. [4]). Equivalently, this can be written as

$$\begin{aligned}
& \min_{\nu} \quad \nu \\
& \text{s. t.} \quad \min_{\delta} \left\{ \sum_i \delta_i (\nu - |r_i|)_+ : \delta \in \{0, 1\}^m, \sum_i \delta_i = m - q + 1 \right\} \geq \lambda\psi(\beta).
\end{aligned}$$

The constraint $\min_{\delta} \{\sum_i \delta_i (\nu - |r_i|)_+ : \delta \in \{0, 1\}^m, \sum_i \delta_i = m - q + 1\} \geq \lambda\psi(\beta)$ can be reformulated exactly using linear programming duality [9, 2] to derive the constraint

$$\left\{ \exists \rho \in \mathbb{R}, \tau \in \mathbb{R}^m \text{ such that } (m - q + 1)\rho - \sum_i \tau_i \geq \lambda\psi(\beta), \rho - \tau_i \leq (\nu - |r_i|)_+ \forall i, \tau \geq 0 \right\}.$$

Note that $\pi_i = (\nu - |r_i|)_+$ can be reformulated exactly as

$$\{\pi_i \geq \nu - |r_i|, \pi_i \geq 0, (\pi_i - \nu - |r_i|, \pi_i) : \text{SOS-1}\}.$$

If we again reformulate $a_i = |r_i|$ exactly as in the nominal formulation using

$$\{a_i = r_i^+ + r_i^-, r_i = r_i^+ - r_i^-, (r_i^+, r_i^-) : \text{SOS-1}\},$$

this leads directly to formulation (13). □

4 On the equivalence of robustification and regularization in matrix estimation problems

A substantial body of problems at the core of modern developments in statistical estimation involves underlying matrix variables. Two prominent examples which we consider here are matrix completion and Principal Component Analysis (PCA). In both cases we show that a common choice of the regularization problem corresponds exactly to a robustification of the nominal problem subject to uncertainty. In doing so we expand the existing knowledge of robustification for vector regression to a novel and substantial domain. We begin by reviewing these two problem classes before introducing a simple model of uncertainty analogous to the vector model of uncertainty.

4.1 Problem classes

In matrix completion problems one is given data $Y_{ij} \in \mathbb{R}$ for $(i, j) \in E \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$. One problem of interest is rank-constrained matrix completion

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{X}\|_{P(F_2)} \\ \text{s. t.} \quad & \text{rank}(\mathbf{X}) \leq k, \end{aligned} \tag{14}$$

where $\|\cdot\|_{P(F_2)}$ denotes the projected 2-Frobenius seminorm, namely,

$$\|\mathbf{Z}\|_{P(F_2)} = \left(\sum_{(i,j) \in E} Z_{ij}^2 \right)^{1/2}.$$

Matrix completion problems appear in a wide variety of areas. One well-known application is in the Netflix challenge [29], where one wishes to predict user movie preferences based on a very limited subset of given user ratings. Here rank-constrained models are important in order to obtain parsimonious descriptions of user preferences in terms of a limited number of significant latent factors [7]. The rank-constrained problem (14) is typically converted to a regularized form with rank replaced by the nuclear norm σ_1 (the sum of singular values) to obtain the convex problem

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{P(F_2)} + \lambda \|\mathbf{X}\|_{\sigma_1}.$$

In what follows we show that this regularized problem can be written as an uncertain version of a nominal problem $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{P(F_2)}$.

Similarly to matrix completion, PCA typically takes the form

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{X}\| \\ \text{s. t.} \quad & \text{rank}(\mathbf{X}) \leq k, \end{aligned} \tag{15}$$

where $\|\cdot\|$ is either the usual Frobenius norm $F_2 = \sigma_2$ or the operator norm σ_∞ , and $\mathbf{Y} \in \mathbb{R}^{m \times n}$. PCA arises naturally by assuming that \mathbf{Y} is observed as some low-rank matrix \mathbf{X} plus noise: $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. The solution to (15) is well-known to be a truncated singular value decomposition which retains the k largest singular values [18]. PCA is popular for a variety of applications where dimensional reduction is desired.

A variant of PCA known as robust PCA [13] operates under the assumption that some entries of \mathbf{Y} may be grossly corrupted. Robust PCA assumes that $\mathbf{Y} = \mathbf{X} + \mathbf{E}$, where \mathbf{X} is low rank and \mathbf{E} is sparse (few nonzero entries). Under this model robust PCA takes the form

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_1} + \lambda \|\mathbf{X}\|_{\sigma_1}. \tag{16}$$

Here again we can interpret $\|\mathbf{X}\|_{\sigma_1}$ as a surrogate penalty for rank. In the spirit of results from compressed sensing on exact ℓ_1 recovery, it is shown in [13] that (16) can exactly recover the true \mathbf{X}_0 and \mathbf{E}_0 assuming that the rank of \mathbf{X}_0 is small and \mathbf{E}_0 is sufficiently sparse. Below we derive explicit expressions for PCA subject to certain types of uncertainty; in doing so we show that robust PCA does not correspond to a robust version of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\sigma_\infty}$ or $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2}$ for *any* model of additive linear uncertainty.

4.2 Models of uncertainty

For these two problem classes we now detail a model of uncertainty. Our underlying problem is of the form $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$, where \mathbf{Y} is given data (possibly with some unknown entries). As with the vector case, we do not concern ourselves with uncertainty in the observed \mathbf{Y} because modeling uncertainty in \mathbf{Y} simply leads to a different choice of loss function. To be precise, if $\mathcal{U} \subseteq \mathbb{R}^{m \times n}$ and g is convex loss function then

$$\bar{g}(\mathbf{Y} - \mathbf{X}) := \max_{\Delta \in \mathcal{U}} g((\mathbf{Y} + \Delta) - \mathbf{X})$$

is a new convex loss function \bar{g} of $\mathbf{Y} - \mathbf{X}$.

As in the vector case we assume a linear model of uncertainty in the measurement of \mathbf{X} :

$$Y_{ij} = X_{ij} + \left(\sum_{\ell k} \Delta_{\ell k}^{(ij)} X_{\ell k} \right) + \epsilon_{ij},$$

where $\Delta^{(ij)} \in \mathbb{R}^{m \times n}$; alternatively, in inner product notation, $Y_{ij} = X_{ij} + \langle \Delta^{(ij)}, \mathbf{X} \rangle + \epsilon_{ij}$. This linear model of uncertainty captures a variety of possible forms of uncertainty and accounts for possible interactions among different entries of the matrix \mathbf{X} . Note that in matrix notation, the nominal problem becomes, subject to linear uncertainty in \mathbf{X} ,

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|,$$

where here \mathcal{U} is some collection of linear maps and $\Delta \in \mathcal{U}$ is defined as $[\Delta(\mathbf{X})]_{ij} = \langle \Delta^{(ij)}, \mathbf{X} \rangle$, where again $\Delta^{(ij)} \in \mathbb{R}^{m \times n}$ (all linear maps can be written in such a form). Note here the direct analogy to the vector case, with the notation $\Delta(\mathbf{X})$ chosen for simplicity. (For clarity, note that Δ is not itself a matrix, although one could interpret it as a matrix in $\Delta^{mn \times mn}$, albeit at a notational cost; we avoid this here.)

We now outline some particular choices for uncertainty sets. As with the vector case, one natural set is an induced uncertainty set. Precisely, if $g, h : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ are functions, then we define an induced uncertainty set

$$\mathcal{U}_{(h,g)} := \{\Delta : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n} \mid \Delta \text{ is a linear map and } g(\Delta(\mathbf{X})) \leq \lambda h(\mathbf{X}) \forall \mathbf{X} \in \mathbb{R}^{m \times n}\}.$$

As before, when g and h are both norms, $\mathcal{U}_{(h,g)}$ is precisely a ball of radius λ in the induced norm

$$\|\Delta\|_{(h,g)} = \max_{\mathbf{X}} \frac{g(\Delta(\mathbf{X}))}{h(\mathbf{X})}.$$

There are also many other possible choices of uncertainty sets. These include the spectral uncertainty sets

$$\mathcal{U}_{\sigma_p} = \{\Delta : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n} \mid \Delta \text{ linear, } \|\Delta\|_{\sigma_p} \leq \lambda\},$$

where we interpret $\|\Delta\|_{\sigma_p}$ as the σ_p norm of Δ in any, and hence all, of its matrix representations. Other uncertainty sets are those such as $\mathcal{U} = \{\Delta : \Delta^{(ij)} \in \mathcal{U}^{(ij)}\}$, where $\mathcal{U}^{(ij)} \subseteq \mathbb{R}^{m \times n}$ are themselves uncertainty sets. These last two models we will not examine in depth here because they are often subsumed by the vector results (note that these two uncertainty sets do not truly involve the matrix structure of \mathbf{X} , and can therefore be “vectorized”, reducing directly to vector results).

4.3 Basic results on equivalence

We now continue with some underlying theorems for our models of uncertainty. As a first step, we provide a proposition on the spectral uncertainty sets. As noted above, this result is exactly Theorem 2, and therefore we will not consider such uncertainty sets for the remainder of the paper.

Proposition 6. For any $q \in [1, \infty]$ and any $\mathbf{Y} \in \mathbb{R}^{m \times n}$,

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{\sigma_q}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|_{F_2} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2} + \lambda \|\mathbf{X}\|_{F_2}.$$

Therefore, for what follows, we restrict our attention to induced uncertainty sets. We begin with an analogous result to Theorem 1. The proof is similar and therefore kept concise. Throughout we always assume without loss of generality that if Y_{ij} is not known then $Y_{ij} = 0$ (i.e., we set it to some arbitrary value).

Theorem 7. If $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a seminorm which is not indentially zero and $h : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a norm, then

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{(h,g)}} g(\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})) = \min_{\mathbf{X}} g(\mathbf{Y} - \mathbf{X}) + \lambda h(\mathbf{X}).$$

Proof. By subadditivity of g we have for any $\Delta \in \mathcal{U} := \mathcal{U}_{(h,g)}$ that

$$g(\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})) \leq g(\mathbf{Y} - \mathbf{X}) + g(\Delta(\mathbf{X})) \leq g(\mathbf{Y} - \mathbf{X}) + \lambda h(\mathbf{X}).$$

Now fix \mathbf{X} . We consider two cases.

1. Suppose $g(\mathbf{Y} - \mathbf{X}) \neq 0$. Pick any $\mathbf{Q} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{Q} \in \operatorname{argmax}_{h^*(\mathbf{A})=1} \langle \mathbf{A}, \mathbf{X} \rangle,$$

where h^* is the dual norm of h . Note in particular that $\langle \mathbf{Q}, \mathbf{X} \rangle = h(\mathbf{X})$. Define the linear map $\hat{\Delta}$ as $\hat{\Delta}(\mathbf{Z}) = \frac{\lambda \langle \mathbf{Q}, \mathbf{Z} \rangle}{g(\mathbf{Y} - \mathbf{X})}(\mathbf{X} - \mathbf{Y})$. Observe that $\hat{\Delta}(\mathbf{X}) = \frac{\lambda h(\mathbf{X})}{g(\mathbf{Y} - \mathbf{X})}(\mathbf{X} - \mathbf{Y})$ and so by the absolute homogeneity of g one has

$$g(\mathbf{Y} - \mathbf{X} - \hat{\Delta}(\mathbf{X})) = \left(1 + \frac{\lambda h(\mathbf{X})}{g(\mathbf{Y} - \mathbf{X})}\right) g(\mathbf{Y} - \mathbf{X}) = g(\mathbf{Y} - \mathbf{X}) + \lambda h(\mathbf{X}).$$

It remains to be shown that $\hat{\Delta} \in \mathcal{U}$. Let $\mathbf{Z} \in \mathbb{R}^{m \times n}$. It is easy to see that showing $g(\hat{\Delta}(\mathbf{Z})) \leq h(\mathbf{Z})$ is equivalent to proving $|\langle \mathbf{Q}, \mathbf{Z} \rangle| \leq h(\mathbf{Z})$. This is true because

$$|\langle \mathbf{Q}, \mathbf{Z} \rangle| \leq \max_{h^*(\mathbf{A})=1} \langle \mathbf{A}, \mathbf{Z} \rangle = h(\mathbf{Z}).$$

Therefore $\hat{\Delta} \in \mathcal{U}$, completing this case.

2. Suppose $g(\mathbf{Y} - \mathbf{X}) = 0$. As in the vector case pick some $\tilde{\mathbf{Z}} \in \mathbb{R}^{m \times n}$ with $g(\tilde{\mathbf{Z}}) = 1$. Define $\hat{\Delta}$ as $\hat{\Delta}(\mathbf{Z}) = \lambda \langle \mathbf{Q}, \mathbf{Z} \rangle \tilde{\mathbf{Z}}$. The rest follows *mutatis mutandis*.

In both cases we have shown that

$$\max_{\Delta \in \mathcal{U}} g(\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})) = g(\mathbf{Y} - \mathbf{X}) + \lambda h(\mathbf{X}),$$

concluding the proof. \square

This theorem leads to an immediate corollary:

Corollary 3. For any norm $\|\cdot\| : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ and any $p \in [1, \infty]$

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{(\sigma_p, \|\cdot\|)}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\| = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\| + \lambda \|\mathbf{X}\|_{\sigma_p}.$$

In the two subsections which follow we study the implications of Theorem 7 for matrix completion and PCA.

4.4 Robust matrix completion

We now proceed to apply Theorem 7 for the case of matrix completion. Note that the projected Frobenius “norm” $P(F_2)$ is a seminorm. Therefore, we arrive at the following corollary:

Corollary 4. For any $p \in [1, \infty]$ one has that

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{(\sigma_p, P(F_2))}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|_{P(F_2)} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{P(F_2)} + \lambda \|\mathbf{X}\|_{\sigma_p}.$$

In particular, for $p = 1$ one exactly recovers so-called nuclear norm penalized matrix completion:

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{P(F_2)} + \lambda \|\mathbf{X}\|_{\sigma_1}.$$

Note that here we have shown that a nuclear norm penalty can arise without directly appealing to sparsity, where sparsity is here defined by rank (nuclear norm is the convex envelope of the rank function on the ball $\{\mathbf{X} : \|\mathbf{X}\|_{\sigma_\infty} \leq 1\}$, which is why nuclear norm is typically used to replace rank [20, 26]). In light of Remark 1, it is not surprising that we can derive a nuclear norm penalty without appealing directly to it; in other words, while the nuclear norm may arguably appear tautologically given the choice of induced uncertainty set $\mathcal{U}_{(\sigma_1, P(F_2))}$, this is not the case.

We detail this argument as before. For any $p \in [1, \infty]$ and $\Gamma = \{\mathbf{X} \in \mathbb{R}^{m \times n} : \|\mathbf{X}\|_{\sigma_p} \leq 1\}$, one can show that

$$\mathcal{U}_{(\sigma_1, P(F_2))} = \left\{ \Delta \text{ linear} : \max_{\mathbf{X} \in \Gamma} \frac{\|\Delta(\mathbf{X})\|_{P(F_2)}}{\text{rank}(\mathbf{X})} \leq \lambda \right\}. \quad (17)$$

Therefore, similar to the vector case with an underlying ℓ_0 penalty which becomes a Lasso ℓ_1 penalty, rank leads to the nuclear norm from the robustification setting without appealing directly to convexity.

4.5 Robust PCA

We now turn our attention to the implications of Theorem 7 for PCA. We begin by noting robust analogues of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$ under the F_2 and σ_∞ norms:

Corollary 5. For any $p \in [1, \infty]$ one has that

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{(\sigma_p, F_2)}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|_{F_2} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2} + \lambda \|\mathbf{X}\|_{\sigma_p}$$

and

$$\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}_{(\sigma_p, \sigma_\infty)}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|_{\sigma_\infty} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\sigma_\infty} + \lambda \|\mathbf{X}\|_{\sigma_p}.$$

We continue by considering robust PCA as presented in [13]. Suppose that \mathcal{U} is some collection of linear maps $\Delta : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ and $\|\cdot\|$ is some norm so that for any $\mathbf{Y}, \mathbf{X} \in \mathbb{R}^{m \times n}$

$$\max_{\Delta \in \mathcal{U}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\| = \|\mathbf{Y} - \mathbf{X}\|_{F_1} + \lambda \|\mathbf{X}\|_{\sigma_1}.$$

It is easy to see that this implies $\|\cdot\| = \|\cdot\|_{F_1}$. These observations, combined with Theorem 7, imply the following:

Proposition 7. The problem (16) can be written as an uncertain version of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$ subject to additive, linear uncertainty in \mathbf{X} if and only if $\|\cdot\|$ is the 1-Frobenius norm. In particular, (16) does not arise as uncertain versions of PCA (using F_2 or σ_∞) under such a model of uncertainty.

This result is not entirely surprising. This is because robust PCA attempts to solve, based on its model of $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ where \mathbf{X} is low-rank and \mathbf{E} is sparse, a problem of the form

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_0} + \lambda \text{rank}(\mathbf{X}),$$

where $\|\mathbf{A}\|_{F_0}$ is the number of nonzero entries of \mathbf{A} . In the usual way, F_0 and rank are replaced with surrogates F_1 and σ_1 , respectively. Hence, (16) appears as a convex, regularized form of the problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{X}\|_{F_1} \\ \text{s. t.} \quad & \text{rank}(\mathbf{X}) \leq k. \end{aligned}$$

Again, as with matrix completion, it is possible to show that (16) and uncertain forms of PCA with a nuclear norm penalty (as appearing in Corollary 5) can be derived using the true choice of penalizer, rank, instead of imposing an *a priori* assumption of a nuclear norm penalty. We summarize this, without proof, as follows:

Proposition 8. For any $p \in [1, \infty]$ and any norm $\|\cdot\|$,

$$\min_{\mathbf{X} \in \Gamma} \max_{\Delta \in \mathcal{U}_{\Gamma(\text{rank}, \|\cdot\|)}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\| = \min_{\mathbf{X} \in \Gamma} \|\mathbf{Y} - \mathbf{X}\| + \lambda \|\mathbf{X}\|_{\sigma_1},$$

where $\Gamma = \{\mathbf{X} \in \mathbb{R}^{m \times n} : \|\mathbf{X}\|_{\sigma_p} \leq 1\}$ and $\mathcal{U}_{\Gamma(\text{rank}, \|\cdot\|)} = \left\{ \Delta \text{ linear} : \max_{\mathbf{X} \in \Gamma} \frac{\|\Delta(\mathbf{X})\|}{\text{rank}(\mathbf{X})} \leq \lambda \right\}.$

4.6 Non-equivalence of robustification and regularization

As with vector regression it is not always the case that robustification is equivalent to regularization in matrix regression problems. For completeness we provide analogues here of the linear regression results. We begin by stating results which follow over with essentially identical proofs from the vector case; proofs are not included here. Then we characterize precisely when another plausible model of uncertainty leads to equivalence.

We begin with the analogue of Proposition 2.

Proposition 9. Let $\mathcal{U} \subseteq \{\text{linear maps } \Delta : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}\}$ be any non-empty, compact set and $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ a seminorm. Then there exists some seminorm $\bar{h} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ so that for any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m \times n}$,

$$\max_{\Delta \in \mathcal{U}} g(\mathbf{Z} + \Delta(\mathbf{X})) \leq g(\mathbf{Z}) + \bar{h}(\mathbf{X}),$$

with equality when $\mathbf{Z} = \mathbf{0}$.

As before with Theorem 4 and Propositions 3 and 4, one can now compute \bar{h} for a variety of problems.

Proposition 10. For any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m \times n}$,

$$\|\mathbf{Z}\|_{F_p} + \frac{\lambda}{\delta_{mn}(q, p)} \|\mathbf{X}\|_{F_{q^*}} \leq \max_{\Delta \in \mathcal{U}_{F_q}} \|\mathbf{Z} + \Delta(\mathbf{X})\|_{F_p} \leq \|\mathbf{Z}\|_{F_p} + \lambda \delta_{mn}(p, q) \|\mathbf{X}\|_{F_{q^*}}, \quad (18)$$

where $\|\Delta\|_{F_q}$ is interpreted as the F_q norm on the matrix representation of Δ in the standard basis. In particular, if $p \neq q$ and $p \in (1, \infty)$, then for any $\mathbf{X} \neq \mathbf{0}$ the upper bound in (18) is strict for almost all \mathbf{Z} (so long as $mn \geq 2$). Further, when $p \neq q$ and $p \in (1, \infty)$, the gap in the lower bound in (18) is arbitrarily small for all \mathbf{X} .

Proposition 11. For any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m \times n}$,

$$\|\mathbf{Z}\|_p + \frac{\lambda}{\delta_{mn}(2, p)} \|\mathbf{X}\|_{F_2} \leq \max_{\Delta \in \mathcal{U}_{\sigma_q}} \|\mathbf{Z} + \Delta(\mathbf{X})\|_{F_p} \leq \|\mathbf{Z}\|_{F_p} + \lambda \delta_{mn}(p, 2) \|\mathbf{X}\|_{F_2}. \quad (19)$$

In particular, if $p \notin \{1, 2, \infty\}$, then for all $\mathbf{X} \neq \mathbf{0}$ the upper bound in (19) is strict for almost all \mathbf{Z} (so long as $mn \geq 2$). Further, if $p \notin \{1, 2, \infty\}$, the gap in the lower bound in (19) is arbitrarily small for all \mathbf{X} .

We now turn our attention to non-equivalencies which may arise under different models of uncertainty instead of the general matrix model of linear uncertainty which we have included here, where

$$[\Delta(\mathbf{X})]_{ij} = \sum_{\ell k} \Delta_{\ell k}^{(ij)} X_{\ell k} = \langle \Delta^{(ij)}, \mathbf{X} \rangle,$$

with $\Delta^{(ij)} \in \mathbb{R}^{m \times n}$. Another plausible model of uncertainty is one for which the j th column of $\Delta(\mathbf{X})$ only depends on \mathbf{X}_j , the j th column of \mathbf{X} (or, for example, with columns replaced by rows). We now examine such a model. In this setup, we now have n matrices $\Delta^{(j)} \in \mathbb{R}^{m \times m}$ and we define the linear map Δ so that the j th column of $\Delta(\mathbf{X}) \in \mathbb{R}^{m \times n}$, denoted $[\Delta(\mathbf{X})]_j$, is $[\Delta(\mathbf{X})]_j := \Delta^{(j)} \mathbf{X}_j$, which is simply matrix vector multiplication. Therefore,

$$\Delta(\mathbf{X}) = [\Delta^{(1)} \mathbf{X}_1 \quad \dots \quad \Delta^{(n)} \mathbf{X}_n]. \quad (20)$$

For an example of where such a model of uncertainty may arise, we consider matrix completion in the context of the Netflix problem. If one treats \mathbf{X}_j as user j 's true ratings, then such a model addresses uncertainty within a given user's ratings, while not allowing uncertainty to have cross-user effects. This model of uncertainty does not rely on true matrix structure and therefore reduces to earlier results on non-equivalence in vector regression. As an example of such a reduction, we state the following proposition characterizing equivalence. Again, this is a direct modification of Theorem 4 and the proof we do not include here.

Proposition 12. For the model of uncertainty in (20) with $\Delta^{(j)} \in \mathcal{U}_{F_{q_j}}$ for $j = 1, \dots, n$, where $q_j \in [1, \infty]$, one has for the problem $\min_{\mathbf{X}} \max_{\Delta \in \mathcal{U}} \|\mathbf{Y} - \mathbf{X} - \Delta(\mathbf{X})\|_{F_p}$ that \bar{h} is defined as

$$\bar{h}(\mathbf{X}) = \lambda \left(\sum_j \delta_m^p(p, q_j) \|\mathbf{X}_j\|_{q_j^*}^p \right)^{1/p}. \quad (21)$$

Further, under such a model of uncertainty, robustification is equivalent to regularization with \bar{h} if and only if $p \in \{1, \infty\}$ or $p = q_j$ for all $j = 1, \dots, n$.

While the case of matrix regression offers a large variety of possible models of uncertainty, we see again as with vector regression that this variety inevitably leads to scenarios in which robustification is no longer directly equivalent to regularization. We summarize the conclusions of this section in Table 4.

Loss function	Uncertainty set	$\bar{h}(\mathbf{X})$	Equivalence if and only if
seminorm g	$\mathcal{U}_{(h,g)}$ (h norm)	$\lambda h(\mathbf{X})$	always
F_p	\mathcal{U}_{σ_q}	$\lambda \delta_{mn}(p, 2) \ \mathbf{X}\ _{F_2}$	$p \in \{1, 2, \infty\}$
F_p	\mathcal{U}_{F_q}	$\lambda \delta_{mn}(p, q) \ \mathbf{X}\ _{F_{q^*}}$	$p = q$ or $p \in \{1, \infty\}$
F_p	\mathcal{U} in (20) with $\Delta^{(j)} \in \mathcal{U}_{F_{q_j}}$	(21)	$(p = q_j \ \forall j)$ or $p \in \{1, \infty\}$

Table 4: Summary of equivalencies for robustification with uncertainty set \mathcal{U} and regularization with penalty \bar{h} , where \bar{h} is as given in Proposition 9. Here by equivalence we mean that for all $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m \times n}$, $\max_{\Delta \in \mathcal{U}} g(\mathbf{Z} + \mathbf{X}) = g(\mathbf{Z}) + \bar{h}(\mathbf{X})$, where g is the loss function, i.e., the upper bound \bar{h} is also a lower bound. Here δ_{mn} is as in Theorem 4. Throughout $p, q \in [1, \infty]$ and $mn \geq 2$.

5 Conclusion

In this work we have considered the robustification of a variety of problems from classical and modern statistical regression as subject to data uncertainty. We have taken care to emphasize that there is a fine line between this process of robustification and the usual process of regularization, and that the two are not always directly equivalent. While deepening this understanding we have also extended this connection to new domains, such as in least quantile regression, matrix completion, and PCA.

Throughout we promulgate that the emphasis on sparsity as a driver for modern, massive-scale statistical estimation is misplaced; instead, we contend that underlying robustness properties are the key to success of many of these modern methods, in particular Lasso and matrix completion, consistent with recent work in [6]. In a world where society broadly relies on noisy data for a myriad of estimation problems across engineering, science, and industry, such robustness properties are not only inherently desirable but necessary in making effective, informed decisions in the face of uncertainty.

References

- [1] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
- [2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
- [3] T. Bernholt. Robust estimators are hard to compute. 2005. Technical Report No. 52/2005, University of Dortmund.
- [4] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, second edition, 1999.
- [5] D. Bertsimas, D.B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.
- [6] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. 2014. submitted.
- [7] D. Bertsimas and R. Mazumder. Factor analysis via a modern optimization lens. 2013. submitted.

- [8] D. Bertsimas and R. Mazumder. Least quantile of squares regression via modern optimization. *Ann. Stats.*, 2014.
- [9] D. Bertsimas and J.N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.
- [10] P. Bonami, M. Kilinc, and J. Linderoth. *Mixed integer nonlinear programming*, chapter Algorithms and software for convex mixed integer nonlinear programs. Springer, 2012.
- [11] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- [12] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure & Appl. Math.*, 59:1207–1223, 2005.
- [13] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *J. ACM*, 58(3):11:1–37, 2011.
- [14] P.L. Combettes and V.R. Wajs. Signal recovering by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–200, 2005.
- [15] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2005.
- [16] D.L. Donoho. Compressed sensing. *IEEE Trans. Inf. Thy.*, 52:1289–1306, 2006.
- [17] D.L. Donoho and P.J. Huber. The notion of breakdown point. *A. Festschrift for Erich L. Lehmann*, pages 157–84, 1983.
- [18] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–8, 1936.
- [19] Y.C. Eldar and G. Kutyniok, editors. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.
- [20] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.
- [21] L. El Ghaoui and H. Lebre. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–64, 1997.
- [22] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.
- [23] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, second edition, 2013.
- [24] H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. *NIPS*, 2007.
- [25] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010.
- [26] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Proc. Forty-Fifth Annual Allerton Conf.*, pages 42–48, 2007.
- [27] R.T. Rockafeller. *Convex analysis*. Princeton University Press, 1970.
- [28] P.J. Rousseeuw. Least median of squares regression. *J. Amer. Stat. Assoc.*, 79:871–80.

- [29] SIGKDD and Netflix. Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Proc. KDD Cup & Workshop*, 2007.
- [30] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc., Ser. B*, 58:267–288, 1996.
- [31] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. *IEEE Trans. Info. Thy.*, 56(7):3561–74, 2010.